DOCUMENT RESUME

ED 189 109 TM 800 271

AUTHOR Lenke, Joanne M.: Canner, Jane M.
TITLE A Design for Linking Tests of a Multi-Level, Multi-Form Achievement Series.
PUB DATE Apr 80
NOTE 7p.: Paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 8-10, 1980).

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Achievement Tests: *Difficulty Level: Elementary Secondary Education: *Equated Scores: *Scaling: Test Interpretation: *Test Items
IDENTIFIERS *Rasch Model: *Test Linking

ABSTRACT

Traditionally, comparable content-area forms and levels of a multi-level, multi-form achievement test series have been equated using the equipercentile method. There is some evidence to indicate, however, that better procedures are needed. A linking procedure is being used to equate simultaneously 40 different test forms of an achievement test: four parallel test forms at each of ten levels of difficulty. The application of this linking procedure provides an efficient means of equating tests and has implications for evaluating the effectiveness of Rasch model procedures for scaling and equating. Tests are equated by a Rasch model strategy: two samples of persons respond to two sets of items; the sets share a common subset of items, which presumably have the same average item difficulty on both tests. (Author/CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. Lenke
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A Design for Linking Tests of a Multi-Level Multi-Form Achievement Series

Joanne M. Lenke
Jane M. Canner
The Psychological Corporation

ED189109

Traditionally, comparable content-area test forms and levels of a multi-level, multi-form achievement series have been equated via the equipercentile method. Alternate forms or adjacent levels are commonly administered either to a common sample of students or, perhaps most often, to matched ability samples. Scores on the two tests are considered equivalent if they give rise to the same percentile rank. However, Slinde and Linn (1977) have indicated that, at least with respect to vertical equating, better procedures are needed. Slinde and Linde (1978) continued to explore this notion by attempting to equate an "easy" test and a "difficult" test with Rasch Model procedures, using each of two samples of people having different abilities. They concluded that "based on the application to one group of ability estimates obtained from a different group, the easy and difficult tests do not seem to provide the equivalent measurements that are necessary for longitudinal research." They do caution, however, that their comparisons may have been more extreme than those apt to be encountered when equating levels of a test over different grades. It should be noted here that the Slinde and Linn Rasch Model equating was evaluated in terms of how well ability estimates derived from different sets of items reproduced the ability scale for different populations of people. However, primary interest in equating test forms and/or levels should be focused on how well ability estimates derived from different sets of items reproduce the ability scale for the same population of people.

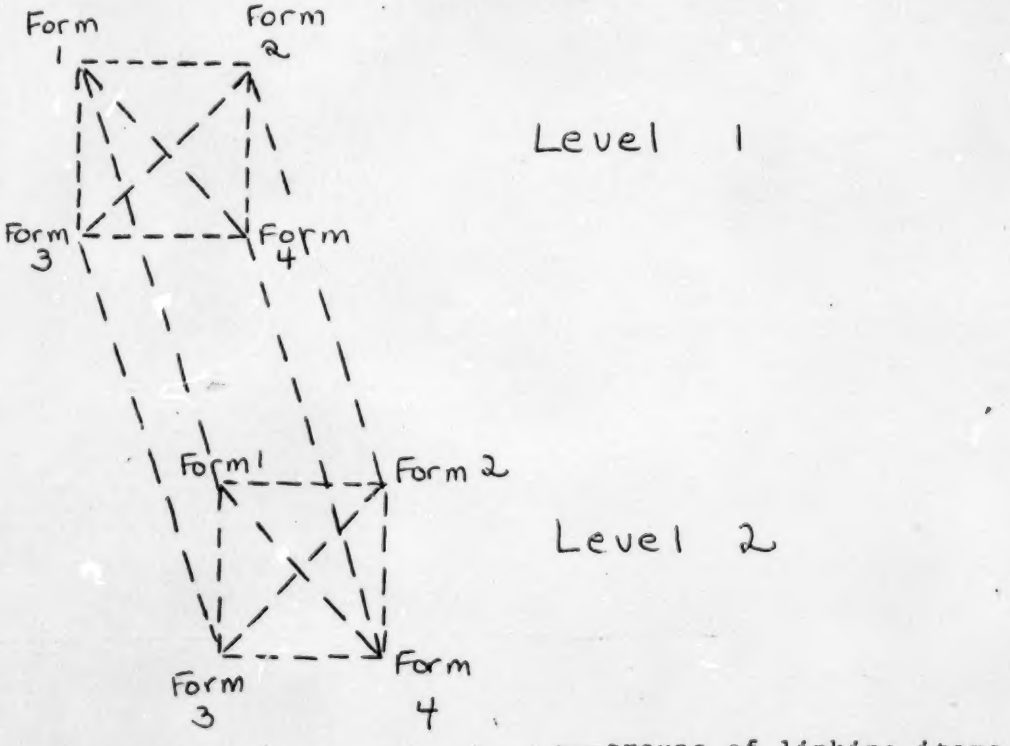
TM 800271

According to Rentz and Bashaw (1977), what must be done in equating two tests is to bring the two ability scales together to form a common scale. Their approach in using the Rasch Model for equating is to estimate the difference between the average item easinesses of the two sets of items and adjust both the item easinesses and the ability values to a single scale having a common origin. They propose two strategies for test equating with the Rasch Model: (1) one sample of persons responds to two sets of items; and (2) two samples of persons respond to two sets of items, with the two sets sharing a common subset of items. Strategy 1 they see as most appropriate for equating alternate test forms; and Strategy 2, as most appropriate for equating test levels differing in difficulty. Since it is this second strategy that is the focus of this paper, additional elaboration follows.

Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston, April, 1980.

The basic assumption underlying the two-sample, two-test, common-subset-of-items approach is that this common subset of items will have the same average item difficulty on both tests. Because the origin for the Rasch scale of any given test is arbitrarily determined by assigning the average item difficulty a Rasch easiness value of zero, the difference between the average item difficulties of the common subset of items serves as an additive constant to adjust the scale of one test to the scale of the other. The purpose of this paper is to describe a procedure used to equate 40 different test forms simultaneously -- four parallel forms at each of ten levels of difficulty.

Wright (1977) suggests that "as the number and difficulty range of items...grows beyond the capacity of any one person (to take the entire set of items), it is necessary to develop an efficient linking system that distributes the items evenly over a network of interlinking tests. The Rasch Model can be used to calibrate all these items on a common scale and to check the various estimated translations against one another for coherence." It is this philosophy that led us to the following linking network design. This diagram represents the linking design for four forms of the test at only two adjacent levels of difficulty; obviously the display of more than two adjacent levels is beyond a three-dimensional configuration. Let it suffice to say that while adjacent-level equating is to be done in a pair-wise fashion, indirect links of non-adjacent levels will serve as checks on the equating constants that have been derived.



Each broken line in the diagram represents a link of common items -- the form-to-form links consist of eight items, and the level-to-level links consist of 16 items. Note also that each form of any one level of the test is also directly linked to each of the other three forms at that level as well as to the corresponding forms of the adjacent higher and lower levels.

The advantage of this particular design is that it provides a means for checking equating constants derived on the basis of direct links, or common sets of items, by a series of indirect confirming links. For example, Form 1 of Level 1 is directly linked to Form 2 of Level 1. But Form 1 is also indirectly linked to Form 2 via a linking path from Form 1 to Form 3 to Form 2, a path from Form 1 to Form 4 to Form 2, and a path from Form 1 to Form 3 to Form 4 to Form 2, and so on, all at Level 1. Other confirming links can be established through the use of other levels.

Let's take a look at an example. Consider a test of mathematics computation, say 40 items, at each of three levels, with four forms at each level. We shall examine a linking design that equates these 12 test "forms" (three levels, four forms each), simultaneously. We'll take a look first at the outcome of such a design with respect to how it will look to a researcher, and then consider some of the problems associated with the test-taking task that may or may not have some effect on the results of equating. Here's how the design works, with a focus on Level 2, the middle level.

Numbers of Items

Basic Form	Level 1				Level 2				Level 3			
	A	B	C	D	A	B	C	D	A	B	C	D
	40	40	40	40	40	40	40	40	40	40	40	40
Form Equating					4(B)	4(A)	4(A)	4(A)				
					4(C)	4(C)	4(B)	4(B)				
					4(D)	4(D)	4(D)	4(C)				
Level Equating					8(1A)	8(1B)	8(1C)	8(1D)				
					8(3A)	8(3B)	8(3C)	8(3D)				
TOTAL (including linking items)					68	68	68	68				

Each of the 28 linking items per form is unique, i.e., an item is used only once in an "overlap" location.

Equating constants, which, in effect, place the items of forms and/or levels on the same Rasch scale of difficulty (and the people taking each on the same Rasch ability scale), can be determined by calculating the mean Rasch item easiness values for pairs of common items and, for each pair, defining the difference between these mean Rasch item easiness values as the equating constant to be applied to each item parameter of the form being equated. Again, to look at our example, let's assume that all items are to be expressed on the Level 2, Form A scale. Equating constants to be applied to the scales of the eleven other test forms can be determined by:

1) direct links between Level 2, Form A and Level 2, Form B, C, and D, each in two directions, i.e., four items from Level 2, Form A to Level 2, Form B and four items from Level 2, Form B to Level 2, Form A; 2) a direct link between Level 2, Form A and Level 1, Form A and between Level 2, Form A and Level 3, Form A, again in two directions, through a total of sixteen level-equating items; and 3) indirect links between Level 2, Form A and Level 1, Forms B, C, and D (through the Level 1, Form A equating) and between Level 2, Form A and Level 3, Forms B, C, and D (through the Level 3, Form A equating).

It is our current thinking that forms within a test level should be equated first, followed by the level equating. The reasons for this should become clear as we examine some of our concerns and ask some very relevant questions about the equating design described above. As we have moved along in the implementation of this design -- assembling test booklets, designing answer sheets, and writing directions -- we have come upon some very practical issues that are worth noting. Perhaps our discussants would like to address themselves to some of these issues.

1. The inclusion of linking items lengthens the "basic form" test significantly. To what extent is this increase in test length likely to affect the parameters of the basic test form and the results of equating, particularly in the primary grades where fatigue and attention span can have significant impact on the test-taking task?

2. The four basic forms at each level were constructed as parallel forms -- parallel in the sense that they measure common sets of objectives. The addition of linking items, then, necessarily implies a certain amount of redundancy of the material assessed. To what extent does this redundancy affect item parameters?

3. Achievement test performance is, by definition, closely related to instruction -- and this is especially true in the primary and elementary school grades. While there is certainly overlap in the objectives measured by adjacent levels of a test, it is not unreasonable

to suggest that both conceptually more difficult material and newly introduced material are assessed on a higher level of the test, and that "easy" objectives, or objectives that have been mastered, tend to "drop out" at the higher level. It was deemed appropriate that for level equating, sixteen unique items per form (eight for the upper level and eight for the lower level) were necessary for linking purposes. On a forty-item test, for example, items have had to be selected very carefully to avoid measuring objectives that have not yet been taught. In this context, can school achievement be considered a continuous ability scale and meet the assumptions of a Rasch Model scale?

4. Different item formats and response modes (booklet versus answer sheet marking and dictated versus non-dictated item stimuli) exist at different test levels. We will have to take a very careful look at these factors as we evaluate the results of equating. The research conducted so far is simply not clear on the extent to which <u>any</u> variation in the presentation of a test item affects its Rasch parameters.

The application of this linking design is a first attempt at a scaling and equating program of considerable magnitude. It not only has potential for an efficient means of equating tests, but also has implications for assessing the effectiveness of Rasch Model scaling and equating procedures that are neither sample dependent nor item dependent. The purpose of our paper was to describe an equating design, and this design is currently being implemented. We hope to have some answers and some further recommendations for you at this time next year.

References

Rentz, R.R. & Bashaw, W.L. The National Reference Scale for Reading: An Application of the Rasch Model. Journal of Educational Measurement, 1977, 14, 161-180.

Slinde, J.A. & Linn, R.A. Vertically Equated Tests: Fact or Phantom? Journal of Educational Measurement, 1977, 14, 23-32.

Slinde, J.A. & Linn, R.A. An Exploration of the Adequacy of the Rasch Model for the Problem of Vertical Equating. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B.D. Solving Measurement Problems with the Rasch Model. Journal of Education Measurement, 1977, 14, 97-116.